Educational equity and teacher discretion effects in high stake exams

Cornelisz, Ilja; Meeter, Martijn; van Klaveren, Chris

published in
Economics of Education Review
2019

DOI (link to publisher)
10.1016/j.econedurev.2019.07.002

document version
Publisher's PDF, also known as Version of record

Economics of Education Review 73 (2019) 101908

Contents lists available at ScienceDirect

Economics of Education Review

journal homepage: www.elsevier.com/locate/econedurev

Educational equity and teacher discretion effects in high stake exams☆

Ilja Cornelisz, Martijn Meeter, Chris van Klaveren*

Amsterdam Center for Learning Analytics (ACLA,acla.amsterdam), Faculty of Behavioral and Movement Sciences, VU University Amsterdam, The Netherlands

ARTICLE INFO

Keywords:
Teacher discretion
Grading
Equity

JEL classification:
I2

ABSTRACT

This study examines teacher discretion effects in Dutch secondary education for the period 2007–2012. Stark discontinuities are observed in the exam grade distribution for high-stakes retaking students and are located at important graduation thresholds. This phenomenon is systematically related to the level of discretion when grading the exam, with results suggesting that approximately 11% of all graduating retakers did so because of teacher discretion. This yields unequal graduation opportunities that are the result of school- and subject choice patterns, since teacher discretion is structurally and selectively exerted at the school-level with the objective to let students on the margin graduate.

1. Introduction

Standardized tests serve to provide objective measures on student performance and these can be high stakes for students as they often determine, at least in part, retention and graduation decisions (Dee, Dobbie, Jacob, & Rockoff, 2016). These standardized tests also have become increasingly central to accountability policies with the objective to evaluate, for example, school and teacher performance. The main intent of test-based accountability policies is to provide incentives that maximize student learning, but perverse incentives resulting from poorly designed accountability policies can have significant, unintended and undesirable consequences (Jacob, 2005). The existing widespread concerns over test validity and the manipulation of scores are therefore not surprising (Dee et al., 2016), yet until recently there has been surprisingly little empirical evidence related to test-based accountability and how it may induce manipulation of student test scores (Jacob, 2005).

Two recent empirical evaluations performed in the United States (Dee et al., 2016) and Sweden (Diamond & Persson, 2016) provide strong evidence that allowing for teacher discretion in grading standardized exams gives all the more reason for policy makers to be concerned. Dee et al. (2016) examine the causes and consequences of test score manipulation of high-stakes exit exams for New York State secondary-school students and find that teachers purposefully moved students just over predefined performance thresholds when grading their own students. Moreover, results varied systematically across and within schools. While black and Hispanic students are more likely to have a test score near a graduation threshold and are thus more often exposed to manipulation, they are less likely to have their test scores manipulated conditional on scoring just below a cutoff. Notably, conditional on scoring near a proficiency cutoff, white and Asian students, students with better baseline scores, and those with good behavioral records are more likely to have their scores manipulated due to teacher discretion. Diamond and Persson (2016) corroborate the existence of test score manipulation for Swedish compulsory schools, and similarly identify 'a bad test day'-effect, suggesting that teachers exploit their discretion to undo potentially harmful consequences of idiosyncratic student performance. In contrast to the results in Dee et al. (2016), their estimates do not suggest that test score manipulation is related to student background characteristics. Furthermore, they find relative homogeneous positive implications for subsequent educational, labor market and life outcomes, highlighting that potential signaling mechanisms resulting from graduation could enhance a student's academic motivation and/or teachers' perception of academic ability.

This study adds to this emerging body of literature on local grading, teacher discretion and test score manipulation (see, also: Burgess & Greaves, 2013; Hanna & Linden, 2012; Lavy, 2008) by evaluating scores on high-stakes standardized exams at the end of secondary education in the Netherlands. It empirically investigates the existence of teacher discretion in grading and potential consequences for inequalities of student graduation opportunities. A specific contribution of this study is the ability to empirically expose the underlying dynamics of teacher discretion mechanisms, by exploiting variation in information, stakes *and* teacher discretion opportunities, as to validate a per-student utility model of teacher discretion.

☆ This authors gratefully acknowledge funding by the Dutch Ministry of Education, Culture and Science (OC&W). We would like to thank Hedvig Horvath, Hessel Oosterbeek for helpful comments. Also, we would like to thank Lianne de Vries for excellent research assistance. All remaining errors are our own.

* Corresponding author.
E-mail addresses: i.cornelisz@vu.nl (I. Cornelisz), c.p.b.j.van.klaveren@vu.nl (C. van Klaveren).

https://doi.org/10.1016/j.econedurev.2019.07.002
Received 14 November 2018; Received in revised form 8 May 2019; Accepted 6 July 2019
Available online 12 July 2019

For this purpose, a unique feature of the Dutch exam system is exploited in that subject teachers grade two similar standardized exams of some of their students twice over a short span of time. Yet, these two attempts differ vastly in terms of the *stakes* at hand and the *information* available, which can impact the validity of the observed student performance measure (Neal, 2013) and the associated teacher grading practices (McMillan & Nash, 2000). For the empirical evaluations, administrative data for the Netherlands is used for the period 2007–2012, covering 99% of the Dutch secondary school exam population. This data is augmented with information on the proportion of open questions on the (retake) exam.[1] Since observed performance gains on the retake exam can potentially be related with teacher discretion, student ability boosting and mean reversion, we use the proportion of open questions as an instrument to identify the effect of teacher discretion. The identifying assumption is that that this measure affects potential teacher discretion (see also Schuurs, Kuhlemeier, & Gitsels, 2017), as it determines the degree of freedom a teacher has to manipulate grades, but not students' ability boosting and mean reversion mechanisms.

First, this paper evaluates if observed performance gains can be predicted by variation in exam openness and to what extent students graduate as a result of teacher discretion. Next, this paper examines whether teacher discretion raises concerns of educational inequity with respect to unequal graduation opportunities within schools with respect to student gender and ethnicity. Finally, the paper explores variation in the proportion of students who graduate by means of a retake exam between schools to establish whether there are also potential concerns of between-school inequity.

This paper proceeds as follows. Section 2, outlines the Dutch institutional background and explains the exam and grading system in detail. Section 3 introduces a so-called graduation game that emerges in this context based on information and stakes and integrates these insights in a theoretical model for teacher discretion in (Dutch exam) grading. Section 4 reports on the data and descriptive statistics. Section 5 shows the empirical findings, and Section 6 summarizes and provides a discussion of the results and their potential policy implications.

2. Dutch education context

2.1. Secondary education in the Netherlands

Upon finishing primary education, children in the Netherlands are tracked into different secondary education levels, with their final track determined after the first of second year of secondary education (Fig. 1). The decision to assign students to a particular track at the start of secondary education is based on both a standardized assessment in taken grade 8 and the advice of the primary school teacher. Three distinct tracks in secondary education can be distinguished. Pre-vocational education (4 years) prepares students for vocational education and comprises 4 separate sub-tracks, secondary general education (5 years) prepares for universities of applied sciences, and pre-university education (6 years) for academic universities. Each track has a matriculation examination in place, in which students take exams in a variety of subjects. This study focuses on students from all three tracks enrolled in the final grade after which they -are expected to- matriculate.

2.2. Secondary school exams and grading system

The school-leaving (matriculation) examination for secondary education in the Netherlands consists, for each subject, of a school examination (*SE*) and a national written examination (hereafter referred to as *C*entral *E*xam, *CE*) at the end of the final school year. Depending on the level of education, students take a CE in roughly 6 to 8 different subjects. The Ministry of Education, Culture and Science prescribes the topics that must be covered in the *SE*s of each subject, but schools have discretion in constructing their own school exams. These school exams usually comprise two or more tests per subject, and can be oral, practical or written. The *CE* for each subject is one test, constructed by the Ministry of Education, Culture and Science, and takes place at a fixed date and time at the end of the final year. The grading scale of each subject is from 1 (lowest) to 10 (highest) and the final grade (grade point average, GPA) for each subject is the arithmetic average of the grades achieved on the school and the central examination (i.e. $GPA_{s(ubject)} = \frac{1}{2}\cdot SE_s + \frac{1}{2}\cdot CE_s$). A GPA of 5.5 is required to pass a particular subject, but since the school leaving examinations consist of 6 or more subjects, there are explicit rules determining whether a student graduates. The specific graduation rules are outlined in Appendix A, but the main determinant for graduation is whether a student has passed (nearly) all individual subjects.

The Dutch examination system give students the opportunity to retake the *CE*, but for *one* subject only. This retake takes place within a week after the first-term results have become known and the highest score on both attempts is used towards determining whether a student has met the requirements for graduation. The formula determining grade point average per subject is then represented by $GPA_{s(ubject)} = \frac{1}{2}\cdot SE_s + \frac{1}{2}\cdot\max(CE_{s1}, CE_{s2})$). Students who fail to graduate will not be eligible to enroll in the post-secondary education sector their track was preparing them for.

Fig. 2 outlines the timing of exam grading activities that are relevant for identifying the possible existence of teacher discretion effects in grading high stakes exams. The figure covers the period from the moment teachers have registered the *SE* grades in a secured digital environment until the moment the grades of the retake exams are publicly announced. Early in May, teachers upload the *SE* grades of their students in *WOLF*, a (web-based) program to exchange exam-related files.[2] Upon successful uploading, teachers can no longer change the *SE* grades registered. At some point mid-May, the CE exams are administered for all subjects and teachers are provided with explicit and strict guidelines regarding the grading procedure. A student's own subject teacher has two weeks to assign a *score* to each answer based on these guidelines and the assigned scores *per question* are uploaded in *WOLF*. Once the scores are uploaded, teachers can no longer change the assigned scores. The Dutch National Institute for Educational Measurement (CITO) has assigned a teacher from a different school (but same subject) to check and re-mark the work (the so-called *second corrector*). The second corrector also has two weeks to review the answers and registers any deviating scores, after which (s)he is redirected to a negotiation page. Also for the second corrector it holds that registered deviations cannot be altered once landed on the negotiation page. Deviations are shown on the negotiation page and both correctors then contact each other to reach agreement. Once agreement has been reached, the second corrector alters the assigned scores and the first corrector must approve that these revised scores are correctly changed in *WOLF*. Upon approval, the assigned scores are final and stored in *WOLF*. In practice, these final scores are very close to the initial scores (i.e. 0.3% lower on average) as uploaded by the first corrector Kuhlemeier and Kremers (2012).[3]

Relevant for the possibility to exert teacher discretion is that CITO

[1] This information is manually obtained from the exam booklets, as uploaded to the ministerial website of the Commission for Tests and Exams (https://www.examenblad.nl/).

[2] WOLF is developed by the Dutch National Institute for Educational Measurement (CITO) charged with all logistics surrounding the national examination in secondary education.

[3] If both correctors cannot reach agreement, school boards from both schools are asked to mediate in the process. If still no consensus is reached, schools can inform the Educational Inspectorate, who can decide to appoint a third independent corrector. In practice, school boards are asked to mediate in only 1% of the cases and the Educational Inspectorate is asked to intervene in only 0.3% of the cases Kuhlemeier and Kremers (2013)

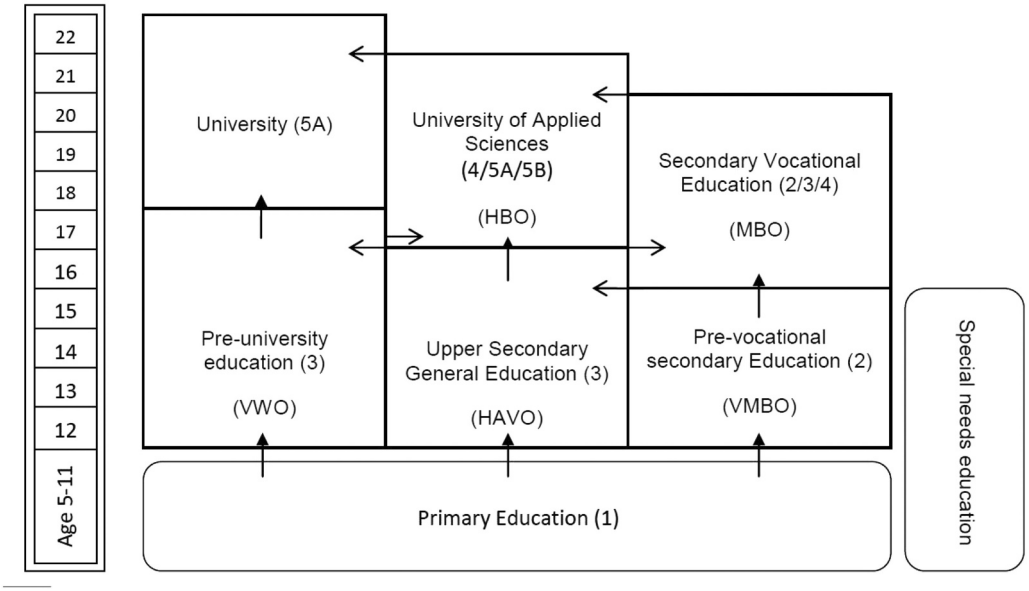


Fig. 1. Dutch education system Note: ISCED levels are shown in parenthesis.

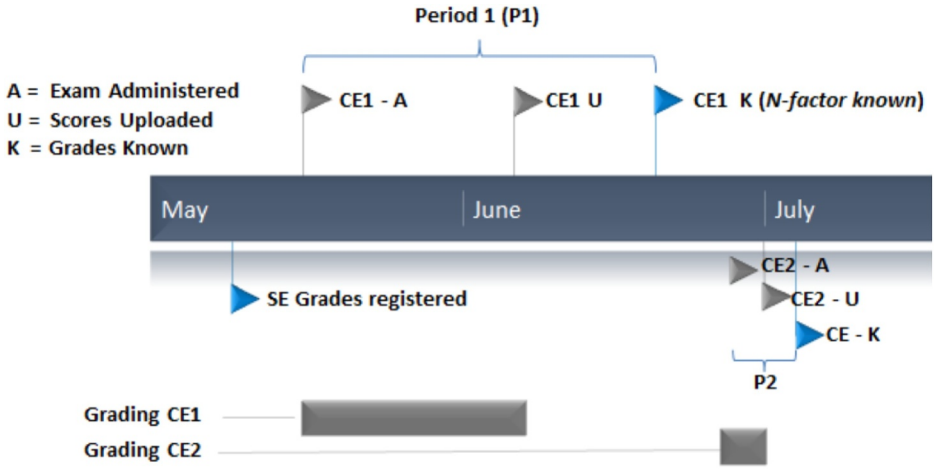


Fig. 2. Timeline of Exam Grading Activities Note: CE1 (CE2) refers to the central exam performance in period 1 (2). SE refers to school-based exams. N-factor is the subject-year-specific term used for converting points to actual grades.

announces the conversion formula required to translate points to an actual grade only after the final scores of all exam students are stored in *WOLF*. This conversion formula contains a subject-specific factor which varies from year to year as to control for erratic differences in the difficulty of a CE exam. In practice, depending on the level of this factor, this can mean a difference in *CE* grade of (over) 2 points on a scale from 1 to 10. Not knowing this factor thus makes that students and teachers cannot have an accurate prior expectation about the number of points required to (just) pass a particular subject. In the conversion formula, this factor is denoted by *N*. This *N*-factor is announced in mid-June, after which students can determine whether they need a retake exam. If a retake exam is required, students will take this exam within a week and the grading procedure will then be similar to the procedure described above. One pivotal difference, given that the *N*-factor is known at the time of grading, is that teachers now know exactly how many points are needed at the retake exam for the student 'at risk' to graduate.

3. The graduation game in the Netherlands

Exams are graded in vastly different contexts in terms of both the stakes at hand and the information available. To be specific, students are allowed to retake *one* exam for one subject which takes place within a week after the results of the first attempt have become known. A retake is often observed if the grade point average (GPA) across all subjects after the first term is insufficient for passing the matriculation examination. The stakes at the retake are thus even higher than in the first-term exams when graduation depends solely on the outcome of this single retake exam. Also, the information to students and teachers between the first and second term is distinctively different. The Dutch Testing Agency (CITO) announces the subject-specific conversion formula used to translate achieved exam points into grades only when the results of the first term are made public. Yet, this same conversion formula then also applies for the subject retake exam that is still yet to be administered and graded. As a result, both teachers and students do not precisely know how many points are needed to pass a particular subject in the first period attempt, but know exactly how many points are needed on the second attempt in order to pass the subject and graduate. We show that when students require a retake exam for graduation, it holds that the optimal strategy of students is to perform as well as they can, and that grade manipulation by means of teacher discretion should explicitly reveal itself in the grading of exams taken in the second term.

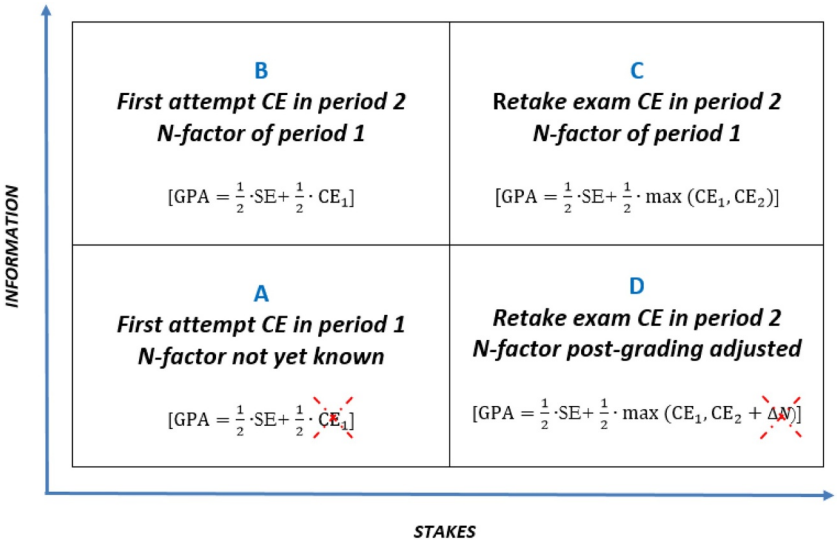


Fig. 3. The Graduation Game .

These different grading environments create a setting we refer to as the graduation game, illustrated in Fig. 3. Distinguishing between these different grading contexts gives insights into when teachers are will exert their discretion in grading an exam as to enable student to -just- graduate. Fig. 3 depicts four possible situations (*A, B, C* and *D*) that occur throughout the exam period and the corresponding *GPA* formula in each situation. This formula represents not only how the *GPA* of the student can be calculated, but also what part of that formula is (un) observed by the teacher *when grading*. In situation *A*, teachers do not observe the N-factor and, therefore, do not observe CE_1 and cannot precisely determine *GPA* when grading. Moreover, teachers are also unaware of the *GPA* for any of the other subjects part of a student's school-leaving examination, making it relatively difficult to pinpoint exactly which students are at risk of not graduating. Yet, in situation B (i.e. a student has a first attempt for a subject in the second term), the situation is markedly different. In this case, teachers and students are aware of the *N*-factor and of the grades achieved through first attempts on other subjects in the first term. Relative to situation A, this additional piece of information enable the subject-teacher in charge of grading the exam to figure out whether a score does or does not yield graduation. This situation somewhat resembles the situation of the retake exam (i.e. situation *C*) in terms of the amount of information available, but the stakes in situation B are lower, because students can still make use of the retake opportunity[4]. Situation C represents the standard retake exam opportunity observed in the second term and, if a student failed to matriculate on the basis of the first term results, both the student and the teacher know that graduation depends solely on the result of this retake exam (i.e. CE_2). One could say that this situation is similar than situation B, but that the urgency (stakes) for exerting teacher discretion and/or manipulating scores as to benefit students is higher. Situation D refers to a specific situation in which the N-factor of the retake exam is adjusted afterwards. Importantly, the N-factor can only be adjusted upwards, such that the initial N-factor is always a correct lower bound of the final grade.

3.1. A Model Of teacher discretion in grading

To gain more insight in the workings of teacher discretion in grading, we integrate the aforementioned graduation game in the model of test score manipulation by Diamond and Persson (2016). The formula used to determine the *CE*-grade can be represented by:[5]

$$CE_i(S_i, N) = (10 - N)\cdot\frac{S_i}{S_{total}} + N, \tag{1}$$

where S_{total} represents the maximum number of points that can be achieved, S_i the actual number of points achieved by student *i*, and *N* the *N*-factor that can take on any value between 0 and 2. The formula highlights that*CE* is determined by both the N-factor and the proportion of points obtained on the exam. In the absence of teacher discretion in grading, Eq. (1) can be rewritten as:

$$CE_i(S_i(SE_i, \varepsilon_i), N) = (10 - N)\cdot\frac{S_i(a_i, \varepsilon_i)}{S_{total}} + N \tag{2}$$

The achieved number of points on the *CE* then depend on student ability, a_i, and an error term, ε_i, which captures the fact that the performance on the central exam may deviate from the true ability of the student. As such, students can have a good test day ($S_i > a_i$) or a bad test day ($S_i < a_i$), reflecting idiosyncratic performance differences with respect to a student's (unobserved) "true" ability (cf. Diamond & Persson (2016)).

Students will only pass a specific subject if $CE \geq 11 - SE$ and teachers can use this information to determine whether they exert discretion in grading (e.g. to ensure a student graduates), thereby awarding additional points to student i, indicated by Δ_i (see Eq. (3)).[6] When we combine this passing rule with Eq. (2) - and rearrange terms-, the threshold of total points (i.e. $S_i + \Delta_i$) required to pass a subject is

[4] Given the incidental nature of situation B, the schedule for this "period3" examinations is determined on a post-hoc basis. These additional retake exams usually take place early August.

[5] We note that in reality the formula is $CE_i = 9/(S_{total}/S_i) + N$. This is a rather inconvenient formula because the *N*-factor can take any value between 0 and 2 such that *CE* may be (lower) higher than the (minimum) maximum *CE* of (1) 10 points when the student answered all exam question (incorrectly) correctly. The examination board, which is a ministerial but independent organization that has the responsibility that the quality and the logistics are guaranteed (see https://www.cvte.nl/) therefore formulated the so-called border-relationships. These border-relationships are grade corrections for when the N-factor is unequal to 1. Information on the exact standardization of the exams can be found at http://wetten.overheid.nl/BWBR0010538/1999-07-07. For the developed model in this section it is important that *CE* is a continuous monotonic increasing function of S_i and *N* and the alternative presented formula for determining *CE* is therefore convenient and without loss of generality.

[6] This can be easily seen, because the subject-specific $GPA = \frac{1}{2}\cdot SE + \frac{1}{2}\cdot CE$ and students pass the subject if $GPA \geq 5.50$.When we substitute $GPA = 5.50$ and rearrange terms we obtain the rule that student pass if $CE \geq 11 - SE$.

given by:

$$S_i(a_i, \varepsilon_i) + \Delta_i = \frac{S_{total}(11 - SE - N)}{10 - N} \tag{3}$$

The left-hand side of Eq. (3) indicates that the total points obtained is the sum of points achieved by the student ($S_i(a_i, \varepsilon_i)$) and any additional points awarded by the teacher by means of exerting discretion (Δ_i). The right-hand side of Eq. (3) highlights that the required total points to pass the exam is conditional on the N-factor and the already registered *SE*-grade. We refer to this right-hand side asκ, or $\kappa(N)$, to illustrate that the teacher knows how many points is required only if the *N*-factor is known.

Based on Eq. (3) and threshold κ, we can then define the following indicator function:

$$t_i = t(a_i, \varepsilon_i, \Delta_i, N) = \begin{Bmatrix} 1 & \text{if } S_i(a_i, \varepsilon_i) + \Delta_i \geq \kappa(N) \\ 0 & \text{Otherwise} \end{Bmatrix}. \tag{4}$$

Eq. (4) reflects that student *i* passes the exam for a certain subject if the obtained points $S_i(a_i, \varepsilon_i) + \Delta_i$ surpasses threshold $\kappa(N)$. A distinct difference between the model of Diamond and Persson (2016) is that our model is not a *fullinformation model* in the sense that teachers do not have full information regarding the threshold value if the *N*-factor has not yet been announced (i.e. situation A). This means that teachers can only effectively target an artificial increase in *CE* grade - exerting their discretion in grading- if the *N*-factor is known. Eq. (4) also captures that teachers will only increase the test score of a student with Δ if this will result in graduation.

Assume that student *i* is taught by teacher *j*. When teacher *j* is manipulating student *i*'s test score effectively, Δ_i is chosen such that the utility function of each student is maximized:

$$u_{ij}(\Delta_i) = \beta_{ij} t_i(SE_i, \varepsilon_i, \Delta_i, N) - c_{ij}(\Delta_i), \quad c_{ij}^{'}(\Delta_i) > 0, \;\; c_{ij}^{''}(\Delta_i) > 0 \tag{5}$$

Parameter β_{ij} reflects teacher*j*'s student-specific desire to raise student *i*'s grade from a fail to a pass, or as Diamond and Persson (2016) remark β_{ij} "*... permits the teacher to use her discretion both in a "corrective" and "discriminatory" fashion*" (p.12). This distinction is important, as the term corrective refers to teachers who may have a preference for compensating a bad test day, while the term discriminatory refers to teachers who may have a preference for increasing test scores of students with certain (background) characteristics. While β_{ij} can vary at the teacher-lever as a result of these student-specific reasons, it also encapsulates any incentives to manipulate that may exist for teachers (e.g. career perspective) and schools (e.g. league tables, accountability) that are irrespective of student *i* and are fixed for teacher *j*. Increasing the points obtained by a student with Δ_i comes at a cost (i.e. $c_{ij}(\Delta_i)$) and these costs are assumed to be strictly increasing and convex. This implies that it becomes increasingly difficult for the teacher to award additional points by means of exerting discretion, given that (1) exams only have a limited set of points that are subjective to teacher discretion, (2) additional points would require rewarding answers that are clearly wrong, and (3) teachers (schools) have to justify their grading results to the second corrector (educational inspectorate). Teachers thus optimally exert their discretion up to the point where the marginal benefits of doing so just offset the marginal costs:

$$\frac{\partial u_{ij}}{\partial \Delta_i} = 0 \Longrightarrow \beta_{ij} \frac{\partial t_i}{\partial \Delta_i} = \frac{\partial c_{ij}}{\partial \Delta_i} > 0 \tag{6}$$

From the model above, a number of empirical hypotheses are derived. Eq. (6) illustrates that teachers will only engage in the costly exertion of teacher discretion if (s)he has a positive (student-specific) desire to do so ($\beta_{ij} > 0$) and when doing so alters the student's subject grade from fail to pass. Furthermore, a teacher would not add points beyond the passing threshold, given that it's costly to do and does not further alter the grade in terms of pass or fail status. As such, if teacher discretion is exerted while grading, this will be targeted effectively and should emerge as a discontinuity in the test score distribution centered around the subject-specific pass-fail threshold. Next, given that knowing exactly the threshold level of points required is contingent on observing the N-factor (i.e. $\kappa(N)$), and given the positive, increasing, convex cost function of adding points by means of teacher discretion ($c_{ij}(\Delta_i) > 0$, $c_{ij}'(\Delta_i) > 0$, $c_{ij}''(\Delta_i) > 0$), test score manipulation will be particularly observed when information available and stakes at hand are both high (i.e. retake attempts of high-risk students). Lastly, the magnitude of points added by the teacher (Δ_i) is conjectured to be increasing in a teacher's student-specific desire (β_{ij}) to engage in this behavior and decreasing in the costs associated with it ($c_{ij}(\Delta_i)$); the latter suggesting that test score manipulation will be more prevalent when the potential to exert teacher discretion is relatively high.

3.2. Empirical strategy for estimating teacher discretion effects

In evaluating the existence of teacher discretion effects for high-stakes retakers in Dutch secondary education, we first display the GPA distributions after the first and second attempt and estimate the size of a discontinuity at the subject-passing threshold of 55 points using McCrary density tests. We do this for different levels of exam openness to see whether the size of such a discontinuity is related to the proportion of open questions. For policy-making it is relevant to get an estimate for the number of high-risk retaking students that are affected by teacher discretion and -moreover- how many graduated by means of teacher discretion.

Thus, in order to estimate teacher discretion effects for high-risk retaking students, we then first determine the observed performance gain as $Gain = CE_2 - CE_1$. Next, we estimate the first-stage discretion-induced performance using a reduced form OLS:

$$Gain_{iklm} = \alpha_0 + \alpha_1 POQ_i + \boldsymbol{X_i}'\boldsymbol{\alpha_2} + \mu_k + \lambda_l + \omega_m + \upsilon_{iklm}, \tag{7}$$

where $Gain_{iklm}$is the observed performance gain on the retake exam for student *i*, observed in educational track *k*, in year *l*, at school *m*. POQ_i is the proportion of open questions for the retake exam of student *i*, X_i a set of student-level characteristics, and we include fixed effects for the educational track (μ_k), year (λ_l) and school (ω_m). Note that we do not observe teacher *j*. υ_{iklm} is assumed to be a random error term, but given that the proportion of open questions varies only per subject-year (e.g. Pre-university track Mathematics in 2007), standard errors are clustered at the subject-year level. The coefficient of interest is α_1 as it depicts the performance gains that can be predicted by variation in proportion open questions.

In estimating to what extent discretion-induced performance gains leads to variation in graduation, a 2SLS procedure is performed in which observed performance gain is instrumented using the proportion of open questions (i.e. Eq. (7)). The second stage then is :

$$Grad_{iklm} = \beta_0 + \beta_1 \hat{Gain}_{iklm} + \boldsymbol{X_i}'\boldsymbol{\beta_2} + \gamma_k + \delta_l + \theta_m + \epsilon_{iklm}, \tag{8}$$

where $Grad_{iklm}$is the observed graduation status after the retake exam for student *i*, observed in educational track *k*, in year *l*, at school *m*. In this second stage equation, $\hat{Gain}_{iklm}$ is the predicted value from the first stage and captures the variation in performance gains that is predicted by the proportion open question questions. X_i is the same set of student-level characteristics, and in similar fashion as the first stage we include fixed effects for the educational track (γ_k), year (δ_l) and school (θ_m). The error term (ϵ_{iklm}) is again clustered at the subject-year level. To avoid the risk of selective sample estimation as a result of potential non-random missing student characteristics, all students are always included in the estimation results, with missing dummy variables for each characteristic included in the model as to account for any level differences between students with and without missing observations on a particular characteristic. The coefficient of interest is β_1 as it indicates the estimate for the discretion-induced graduation effect.

In the final stage of the analysis -and using the estimate of the α_1 coefficient from estimating Eq. (7) we simulate the performance gains for graduating high-stakes retaking student i that would have occurred if no teacher discretion was present and then calculate a counterfactual graduation status for all students, using the final scores thus obtained and the pass-fail regulations appropriate for that specific cohort (see Appendix A). By comparing the number of high-stakes retaking students that actually graduated with the simulated number that would have occurred if no teacher discretion was present gives us a policy-relevant indicator for how many students graduated by means of teacher discretion. To check for potential heterogeneities, we perform the aforementioned estimation strategy for different subgroups of students (i.e. by gender and ethnicity).

4. Data and descriptive statistics

This study uses student-level administrative data on 1.12 million students who are in their final secondary school year in the period 2007–2012. The data contains information on students enrolled in publicly-funded schools, covering 99% of the exam student population. For each student, a list of background characteristics is known, together with the results on school examinations and central exams (for all subjects and both terms). Information about the *N*-factor was derived from the ministerial website of the Commission for Tests and Exams (https://www.examenblad.nl/). The average N-factor in period 1 was 0.95 (SD = 0.45) and the N-term adjustment in the retake was 0.29 (SD = 0.19).

Table 1 compares student characteristics between the full student population and the population of students who make use of their retake opportunity. Among the population of retakers, students who required a retake in order to graduate are labeled '*high-stakes*'. The average student is around 16 years old and has achieved a school exam grade of 6.52, on a 1 (lowest) to 10 (highest) scale. The proportion of students with a migrant background or living in an impoverished neighborhood is, respectively, 0.20 and 0.13. The proportions related to education- level show that most students are enrolled in pre-vocational education (55%) and the least students are enrolled in pre-university education (17%). Students who used their retake opportunity have, on average, lower school exam grades, are somewhat older and more frequently have a migrant background or live in an impoverished neighborhood. Also they are more frequently enrolled in upper general or pre-university education. Whether students required a retake in order to graduate (i.e. the retake is high stakes) is reflected in the lower achieved *SE*-grade of 5.74. Even though these students have scored a lower grade on their school examination, they have otherwise rather similar characteristics to other retakers, except that upper-general education is relatively over-represented.

Table 1
Comparing characteristics of full population with retakers .

	Full Population		Retake Population			
			All		High Stakes	
	Mean	SD	Mean	SD	Mean	SD
Male	0.50	0.50	0.47	0.50	0.48	0.50
Age on October 1st	16.08	0.96	16.25	1.00	16.37	1.02
Non-Dutch Background	0.20	0.40	0.30	0.46	0.32	0.46
Impoverished Neighborhood	0.13	0.34	0.18	0.39	0.19	0.40
SE	6.52	0.84	6.04	0.86	5.74	0.69
Pre-vocational education	0.55	0.50	0.48	0.52	0.43	0.50
Upper general education	0.26	0.44	0.29	0.45	0.34	0.47
Pre-university education	0.17	0.38	0.23	0.42	0.23	0.43
Missing covariate(s)	0.16	0.37	0.20	0.40	0.22	0.41
POQ	0.68	0.29	0.70	0.30	0.72	0.29
CE_1	62.5	11.9	48.8	10.3	45.3	8.38
CE_2	–	–	55.6	13.4	54.4	13.1
N	1,118,650		253,796		136,638	

5. Findings

5.1. Performance gains for marginal students in retake exams

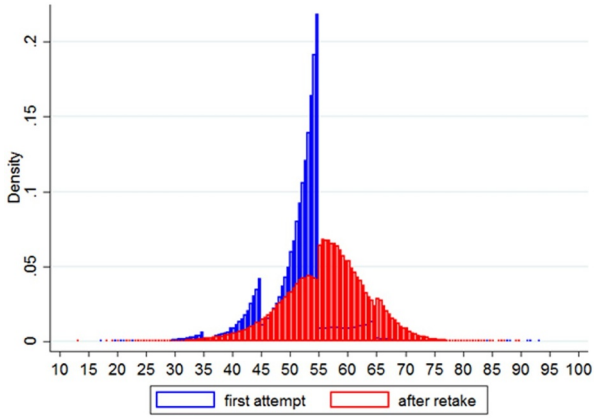
The theoretical model predicts that teachers will (primarily) exert their discretion when the information available and stakes at hand are both high and if doing so would lead to graduation. In total, 253,796 (22.7%) students make use of their retake opportunity, of which 136,638 (53.8%) are required a retake in order to graduate (i.e. high-stakes retakers). Fig. 6 shows for high-stakes retakers the final grades distributions based on only the first attempt (i.e.CE_1) and based on the highest achieved grade achieved in the first attempt and the retake (i.e.max (CE_1, CE_2)). Of these high-stakes retakers, 108,972 students (79.8%) experienced a GPA gain by means of the retake exam and 69,280 students (50.7%) graduate as a result of this GPA gain. The figure shows a large and significant discontinuity at the passing threshold of 5.5, indicating that a substantial fraction of students is transferred from the left to the right of the passing threshold. There are three reasons that could cause such a transfer:

1. Ability boosting, in that students put in a lot of effort in preparing for the retake as to improve their performance,
2. Mean reversion, in that students performed relatively low on the initial attempt ($S_i < a_i$),
3. Teacher discretion, in that teachers exert their discretion to award additional points on the retake exam ($\Delta_i > 0$),

Ability boosting is the process in which students (temporarily) raise their performance level with the objective to graduate. The optimal strategy for high-stakes students is to boost their ability and score as many points as possible on the retake exam, thereby maximizing the probability of graduating and minimizing the uncertainty with respect to whether the scored points are indeed enough for graduation. Since students are able to convert scored points on the retake exam to grades -as the *N*-factor is available to them- it can be argued that it is optimal to achieve only exactly the required number of points needed for graduation, since this maximizes graduation at minimal effort. However, students do not know how much effort is required and cannot exactly convert points on the retake exam to grades when the proportion of open questions is high, which further exemplifies that it is optimal for students to boost their ability as much as possible and to score as many points as possible on the retake exam (Fig. 4, Tables B.2 and A1).

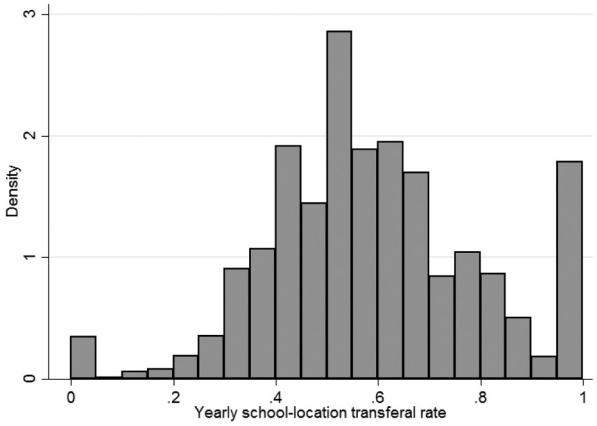
Mean reversion departs from the recognition that the grade of a student is drawn from his/her own (normal) grade distribution, and when the first attempt produced a grade low in this distribution ($S_i < a_i$), the probability that the retake produces a higher grade than the grade achieved in the first attempt is relatively high. Mean reversion can be considered a 'bad test day'-effect and cannot be structurally related to the proportion open questions on the exam.

Finally, the theoretical model of teacher discretion in grading (Section 3) shows that teacher manipulation always results in an improved grade (i.e. $\Delta_i > 0$) and is applied only when when it leads to graduation (i.e. $t_i > 0$). Moreover, the model shows that positive grade manipulation comes at a cost, because exams have a restricted set of open questions that are subjective to teacher discretion. When all exam questions consist of multiple choice questions, then there is no set of points at all that is subjective to teacher discretion. Instead, when there are many open questions, teachers have a larger set of questions to exploit their discretion and increase the grade. Thus, when teachers structurally manipulate exam grades in the retake period, the proportion of open questions of the retake exam should correlate positively with both grade gains and the propensity to graduate.



N= 136698
Discontinuity= 0.44***
SE = 0.017

Fig. 4. Grade distributions based on first and retake attempts.



N = 1176 school locations
Mean = 58.8%
SD = 0.21

Fig. 6. School-level transferal rate distribution.

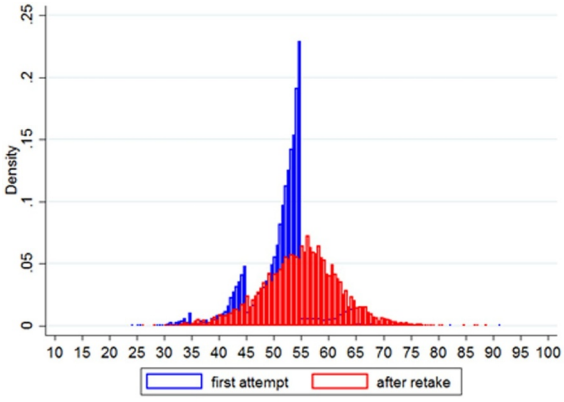
Fig. 5 illustrates the relationship between the proportion of open questions and GPA gains that students achieve by doing a retake exam. The left panel shows the GPA gains when the proportion of open questions is less than 20%, while the right panel shows the GPA gains when the proportion of open questions is more than 80%. Both panels show that students are transferred to the right of the passing threshold, but McCrary density tests (McCrary, 2008) show that the observed discontinuity at the passing threshold (i.e. $GPA = 55$) is much larger when the proportion of open questions is larger (i.e. 0.55 versus 0.14).

These results are indicative that the proportion of open questions on the retake exam can be important in estimating potential teacher discretion effects when comparing high-stakes retaking students. Yet, before presenting the results of the proposed estimation strategy exploiting variation in the proportion open questions on the retake exam, it is relevant to note that no significant discontinuity is observed (0.02) for first attempt scores observed in the first attempt but that a discontinuity *is* observed (0.23) for first-attempt scores for the very small (N = 6,679) and non-random group of students for which a first attempt is observed in the second period (i.e. group B, when the N-factor is known). Despite the selective nature of this subpopulation, finding this discrepancy further corroborates the potential existence of teacher discretion effects. Appendix C elaborates on these findings.

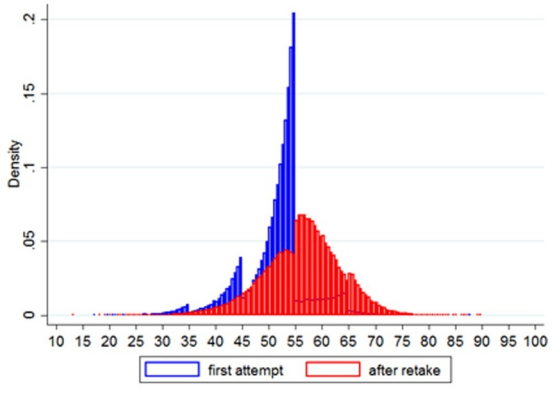
5.2. Teacher discretion effects on student graduation

Based on the aforementioned reasoning, the proportion open questions (POQ) is used as an instrument for identifying teacher discretion effects. In Appendix B, it is shown that for the 386 subject-year clusters considered in the sample of graduating retakers, the proportion of open questions (Fig. B.1) is negatively skewed (M = 0.72 and SD = 0.28). Table 2 Model 1 indicates that POQ does not predict first-attempt performance on the central exam (CE_1), thereby supporting its validity as an instrument to identify teacher discretion effects on the retake exam (i.e.CE_2).

Model 2 in Table 2 considers observed gains on the retake exam (i.e. Eq. (7)) and shows that the estimated coefficient for *POQ* now is 3.6 points and highly significant. In order to isolate variation in retake exam grades resulting from teacher discretion (and not mean reversion or ability boosting), CE_2 is instrumented by proportion open questions



N= 7,471 students
Discontinuity= 0.14***
SE = 0.054



N = 77,721 students
Discontinuity = 0.55***
SE = 0.017

Fig. 5. Grade distributions based on first and second attempt - by exam question openness.

Table 2
Exam openness: CE_1, performance gains and graduation effects.

Outcome	CE_1		CE_2 - CE_1		*Graduation*	
Estimation	*OLS*		*OLS*		*IV.2SLS*	
	1		2		3	
	Coeff	SE	Coeff	SE	Coeff	SE
POQ	0.100	(0.788)	3.593***	(0.845)		
CE_2					0.027***	(0.005)
Constant	41.11***	(2.524)	25.47***	(2.333)	-0.531***	(0.187)
Student-level controls	√		√		√	
Level + Year dummies	√		√		√	
School Location FE	√		√		√	
N	136698		136698		136553	
R^2	0.069		0.241		0.158	
# subject-year clusters	386		386		386	

*Note:*Robust standard errors in parentheses. */**/*** denote significance at a 10/5/1 percent confidence level.

(*POQ*). Doing so reveals that the instrument does not seem to weak with an F-statistic of the empty-model first stage of 25.47 and Table B.1 (Appendix B) gives the first stage results of the full model including all covariates. Model 3 in Table 2 displays the corresponding 2SLS results for the effect of teacher discretion on graduation for retakers and confirms that this effect is positive and statistically significant. This confirms that teachers exploit their discretion to let students graduate who would otherwise not have graduated.

To get an idea of the number of students this concerns, the observed central exam improvement gains for the 69,280 individual graduating retaking students are "corrected" for teacher discretion, using the proportion of open questions on the retake exam and the 95% confidence interval of the POQ parameter estimate obtained by estimating the regression equation of Model 2 for this subset of students. The estimated coefficient of POQ for theses students is 4.21 with 95% CI [2.94, 5.47]. Using these simulated lower and upper bounds of performance gains in the case of no teacher discretion, subject-specific final GPA scores are re-calculated and for each student it is determined whether or not (s)he would have still passed the pivotal graduating threshold. The result of simulation exercise returns a range of 7.2–13.7% (with a mean of 11.0 percent). This suggests that approximately 11% of all students who graduated by means of a retake did so because of teacher discretion. On a yearly basis, this translates to roughly 1300 students who are transferred to graduation as a result of a teacher exerting discretion while grading

5.3. Teacher discretion and unequal graduation opportunities

Notwithstanding that teacher discretion effects can originate from a genuine desire to help students graduate, it can have undesirable effects in that it can cause inequitable within- and between-school variation in graduation opportunities for high-stakes retakers.

With respect to potential concerns of within-school inequity, Table 3 Model 1 again shows the reduced form OLS results for performance gains, but now also disaggregated by gender and Dutch background. The subgroup-specific coefficient estimate of the impact of proportion open questions on performance gains (α_1) is then used to simulate how many graduating high stakes retakings students for each subgroup would have graduated if no teacher discretion would be present. The results reveal that relatively more high-stakes retaking girls benefited from teacher discretion with respect to observed graduation status as well as students with a Dutch background. These differences are the joint result of mechanisms operating both at the teacher-and school-level, which cannot be disentangled as no teacher characteristics are observed.

These results indicate that teacher discretion effects can result in unequal graduation opportunities (by school choice) and that these effects arise because (teachers in) some schools exploit discretion in grading retake exams more than (teachers in) other schools.

With respect to potential concerns of between-school inequity, the proportion of high-stakes retakers at a school that graduated due to the retake exam result -referred to here as the transferal rate- is exploratively examined. The average transferal rate for the sample of high-risk retakers is 54.8%. Figure 8 displays the transferal rate *by school location and year* and the distribution shows a substantial amount of variance, indicating that in some schools no high-stakes retakers are transferred towards graduation in a given year, while in other schools all high-stakes retakers are transferred towards graduation in that year.

Observing variation around the mean is not necessarily problematic with respect to between-school inequity, as long as schools are not structurally located high (or low) in this distribution over time. To examine whether this is the case, we estimate several random effects models shown in Table 4. Baseline model 1 shows that the (weighted) average school-level transferal rate is 58.1% and the intra-class correlation coefficient (ρ) indicates that 65.2% of the observed variation in transferal rate is due to between school variation. When education level dummies are included in model 2, the residual intra-class correlation coefficient becomes lower, but is still 58.6%. It can be argued there may also be sorting effects of students into schools, in that high-quality schools attract better students. However, the random effect analysis is performed only for high-stakes retakers, thus who all required a retake exam for graduation after taking a nationwide standardized exam. Yet, to control for potential sorting effects, student controls are added in model 3, including the school examination grade and the central exam grade of the first period, and the results show that still 56.4% of the variation in school-level transferal rate is between school-level variation. The final model shows that between-school variation is not driven by structural differences between schools in the proportion of open questions of retake exams observed across school locations. These results show that high-stakes retakers are transferred towards graduation structurally more often in some schools than in others. As such, this offers a substantial source of between-school inequity in graduation opportunities for these high-stakes retakers. Importantly, the bottom row in Table 4 shows the results when estimating the same models for first attempt scores instead. For this outcome, between school variation can account for only 2–4% of the total variation instead. The fact that structural differences between schools emerge only after retakes provides further suggestive evidence that this phenomenon is due to differences in teacher discretion effects.

6. Discussion

This study shows that teachers structurally use their discretion to increase the performance of their students with the objective to let them graduate. This discretion is targeted at student who find themselves just below the passing threshold and effectuated when the stakes are highest and there is full information on how to convert assigned points to grades. To distinguish teacher discretion effects from the effects of

Table 3
Exam openness: heterogeneous treatment intensity.

	All		*Boys*		*Girls*		*Dutch*		*non-Dutch*	
Outcome	CE_2 - CE_1		CE_2 - CE_1		CE_2 - CE_1		CE_2 - CE_1		CE_2 - CE_1	
Estimation	*OLS*		*OLS*		*OLS*		*OLS*		*OLS*	
	1		**2**		**3**		**4**		**5**	
	Coeff	SE	Coeff	SE	Coeff	SE	Coeff	SE	Coeff	SE
POQ	3.593***	(0.845)	3.199***	(0.970)	3.293***	(0.970)	3.827***	(0.970)	2.857***	(0.893)
CE_1	−0.533***	(0.016)	−0.559***	(0.017)	−0.559***	(0.017)	−0.532***	(0.017)	−0.542***	(0.017)
SE	0.277***	(0.016)	0.277***	(0.017)	0.271***	(0.020)	0.287***	(0.020)	0.265***	(0.017)
Constant	25.47***	(2.333)	26.48***	(2.645)	30.54***	(3.124)	23.45**	(2.687)	28.46**	(2.799)
Student-level controls	√		√		√		√		√	
Level + Year dummies	√		√		√		√		√	
School Location FE	√		√		√		√		√	
Simulated Graduation Effect	11.0%		10.2%		12.4%		11.2%		9.5%	
N	136698		51002		55608		93760		42938	
R^2	0.241		0.275		0.269		0.223		0.289	
# subject-year clusters	389		323		330		330		330	

*Note:*Robust standard errors in parentheses. */**/*** denote significance at a 10/5/1 percent confidence level.

Table 4
Between-school inequity: random-effects model.

Dependent variable: School-level year transferal rate

	1		2		3		4	
	Coeff	SE	Coeff	SE	Coeff	SE	Coeff	SE
Constant	0.581***	(0.004)	0.573***	(0.004)	0.534***	(0.009)	0.520***	(0.009)
Education level dummies	*No*		*Yes*		*Yes*		*Yes*	
Student controls	*No*		*No*		*Yes*		*Yes*	
Exam openness	*No*		*No*		*No*		*Yes*	
# students	136,698		136,698		136,698		136,698	
# school locations	1,176		1,176		1,176		1,176	
sigma_u	0.143		0.125		0.119		0.119	
sigma_e	0.105		0.105		0.104		0.104	
R^2	0.000		0.098		0.070		0.069	
$\rho_{transferalrate}$	0.652		0.586		0.564		0.564	
$\rho_{firstattemptscore}$	0.041		0.040		0.030		0.023	

Note: */**/*** denote significance at 10/5/1% level (two-sided). Outcome variable in Models 1–4 is the observed yearly transferal rate for the school location a student attends. Standard errors are clustered at the subject-year level. Student-level controls are: boy, age, non-Dutch background, impoverished neighborhood, SE-grade and a dummies for whether a covariate is missing. Education level dummies are: pre-vocational education, upper general education. Exam openness is the percentage of open questions of the retake exam. School-location random effects are based on 1176 school locations.

Table B.1
First-stage results of instrumenting CE_2 scores by Proportion Open Questions for marginal retakers.

Outcome	CE_2	
Estimation	*OLS*	
	1	
	Coeff	SE
Prop. Open Questions (POQ)	3.597***	(0.846)
Constant	25.28***	(2.298)
Student-level controls	√	
Education Level dummies	√	
Year dummies	√	
School Location Fixed Effects	√	
N	136,553	
Adj.R^2	0.243	
# subject-year clusters	389	

Note: Robust standard errors in parentheses. */**/*** denote significance at a 10/5/1 percent confidence level. Outcome variable in Models 1–4 is the observed first-attempt CE-grade for a graduating retaker. Standard errors are clustered at the subject-year level. Student-level controls are: boy, age, non-Dutch background, impoverished neighborhood, SE-grade and a dummies for whether a covariate is missing. Education level dummies are: pre-vocational education, upper general education.

ability exploitation and mean reversion we use the proportion of open questions (summary, essay) as an instrument to identify teacher discretion effects. The identifying assumption is that the proportion open questions is unrelated with ability boosting of students (i.e. students -temporarily- raise their performance level with the objective to graduate) and mean reversion, but can be positively related to teacher discretion. We find that teacher discretion is revealed when the teacher has full information regarding the conversion of points obtained to grades, that the effect is larger in magnitude when students are at risk of not graduating, and larger when teachers have more discretion when grading the exam. The results suggest that approximately 11% of all students who graduated by means of a retake exam did so because of teacher discretion. This roughly translates to 1300 students in any given exam year. This result is derived only from teacher discretion at retake exams and -given that teachers locally grade many more tests (e.g. school exams)- could therefore be considered to be a lower-bound estimate of the overall implications of teacher discretion effects in Dutch secondary education.

Notwithstanding the good intentions teachers arguably have to artificially improve the grades of students who are on the margin of graduation, it may have undesirable inequity effects. First of all, it

Table B.2
Validity instrument: Exam openness and CE_1 performance.

Outcome	CE_1		CE_1		CE_1		CE_1	
Estimation	*OLS*		*OLS*		*OLS*		*OLS*	
	1		2		3		4	
	Coeff	*SE*	*Coeff*	*SE*	*Coeff*	*SE*	*Coeff*	*SE*
POQ	−0.295	(0.640)	−0.064	(0.741)	−0.001	(0.840)	0.100	(0.788)
SE			0.114***	(0.025)	0.115***	(0.028)	0.122***	(0.026)
Constant	44.95***	(0.447)	41.59***	(5.638)	41.15***	(3.116)	41.11***	(2.524)
Student-level controls			√		√		√	
Level + Year dummies					√		√	
School Location FE							√	
N	136698		136698		136698		136698	
R^2	0.016		0.036		0.045		0.069	
# subject-year clusters	389		389		389		389	

Note: Robust standard errors in parentheses. */**/*** denote significance at a 10/5/1 percent confidence level.

Table A1
Selective participation in first exam attempt.

	First Attempt in			
	Term 1		Term 2	
	Mean	SD	Mean	SD
Male	0.50	0.50	0.50	0.50
Age on October 1st	16.07	0.96	16.18	1.05
Migrant Background	0.20	0.40	0.28	0.45
Impoverished Neighborhood	0.13	0.34	0.18	0.39
SE	6.52	0.84	6.36	0.91
Pre-vocational education	0.55	0.50	0.50	0.50
Upper general education	0.26	0.44	0.29	0.46
Pre-university education	0.19	0.39	0.21	0.41
N	1,111,971		6,679	

results in between-subject variation (i.e. the retake subject choice matters for graduation). Secondly, it can cause both inequitable between-school variation (i.e. school-location differences in transferal rates) and within-school variation due to heterogeneity of teacher discretion effects with respect to student-level characteristics. The structural differences observed between schools indicate that teacher discretion effects result in unequal graduation opportunities (by school choice) and these effects arise because (teachers in) some schools exploit their discretion more than (teachers in) other schools.

When objective skills-assessment is a priority of the school-leaving exams, the results presented here show that teacher discretion issues can -at least in theory- be easily resolved by either avoiding teacher discretion when grading, or by imposing that the nation-wide central exams are graded anonymously. Obviously, this can potentially start a different public debate about whether it is desirable that students who are just below the passing threshold have to redo the examination year (either partially or entirely). However, these valid questions are directly targeted at the functioning of the exam system, and stand alone in the fundamental argument that students should have equal educational opportunities.

Appendix A. Graduation rules: Pass-Fail Regulations

For the evaluation window considered in this study, the rules state (REF) that students in pre-vocational education pass the school-leaving examinations if one of the following situations hold:

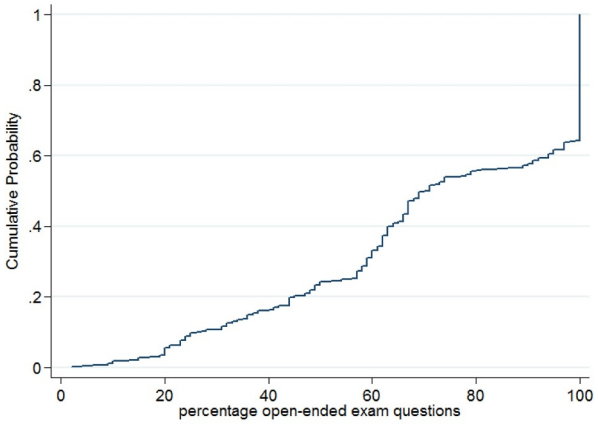
1. GPA for all subjects is at least 5.5
2. GPA for one subject is between 4.5 and 5.45, for all other subjects at least 5.5
3. GPA for one subject is between 3.5 and 4.45, for one subject at least 6.5, and for all other subjects at least 5.5
4. GPA for two subjects is between 4.5 and 5.45, for one subject at least 6.5, and for all other subjects at least 5.5

For students who are in secondary general education, and pre-university education these rules are:

1. GPA for all subjects is at least 5.5
2. GPA for one subject is between 4.5 and 5.45, for all other subjects at least 5.5
3. GPA for one subject is between 3.5 and 4.45 and for all other subjects at least 5.5 and overall GPA is at least 6.0 based on subject grades rounded to whole integers.
4. GPA for two subjects is between 4.5 and 5.45, and for all other subjects at least 5.5 and overall GPA is at least 6.0 based on subject grades rounded to whole integers.
5. GPA for one subject is between 3.5 and 4.45, for one subject between 4.5 and 5.45, and for all other subjects at least 5.5 and overall GPA is at least 6.0 based on subject grades rounded to whole integers.

For 2012, an additional graduation requirement for students in secondary upper general and pre-university education is that the average CE-grade across all subjects is at least 5.5.

Appendix B. Proportion open questions



N = 386 subject-year clusters
Mean = 71.7%
SD = 0.28

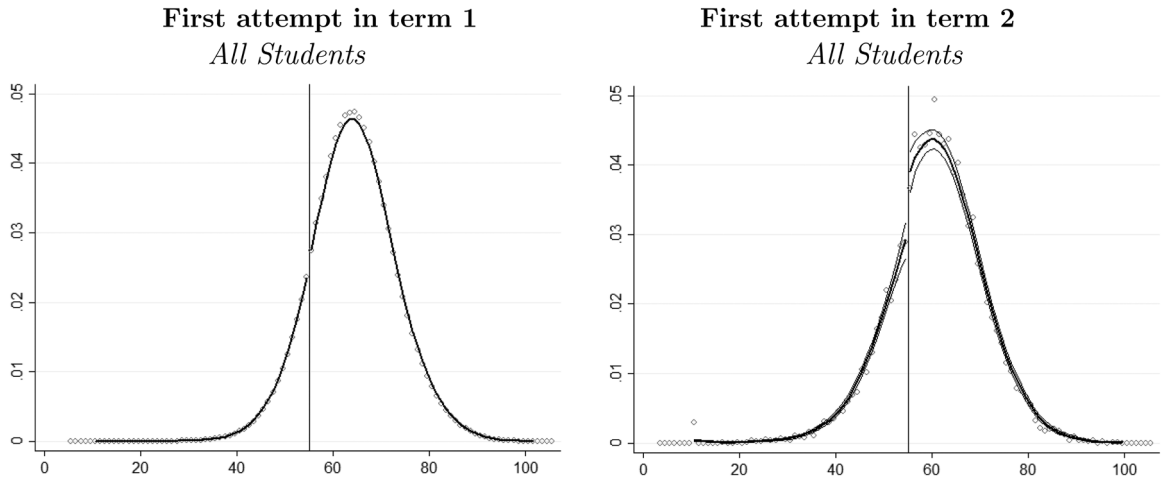
Fig. B.1. Cumulative distribution function of proportion open questions.

Appendix C. Teacher discretion effects and selective participation in first exam attempts

For the vast majority of students (N = 1,111,971), the first attempts are observed in the first term (i.e. situation A), when the *N*-factor required for points-to-grade conversion is unknown. First attempts in the second term (i.e. situation B) are arguably incidental and random (e.g. sickness), such that average student characteristics related to exam performance should not correlate with the incidence of observing a first attempt in the second term. Although this situation indeed occurs for a relatively small group of students (N = 6,679), it is an interesting situation because in term 2 the *N*-factor is known to both teachers and students such that exam scores can be precisely converted to a grade. Furthermore, mean reversion cannot confound the interpretation when comparing first-attempt subject-exam grade distributions across situations A and B. Lastly, we can compare first attempt results for the same group of students, thereby exploiting variation in whether the *N*-factor is known (i.e. situation A versus B), variation in stakes at hand (i.e. all students versus high-stakes retakers), and variation in teacher discretion (i.e. all subjects versus subjects with at least 80% open questions).

Table 3 compares characteristics of students who had their first exam attempt in term 1 with those who had their first attempt in term 2. The table indicates that sample of students withfirst attempt in term 2 (situation B) are not a random sample of the total student population. Students with a first attempt in term 2 are, on average, somewhat older, have achieved lower school exam grades, more frequently have a migrant background and live in an impoverished neighborhood. These differences indicate that a selective group of relatively lower performing students had their first attempt in term 2.

Fig. A1 then compares the subject-specific GPA distributions of first exam attempts for term 1 and 2 separately. The left panel shows no sizable discontinuity (i.e. 0.02) at the passing threshold of 55 (the statistical significance is primarily the result of the large number of observations in



N = 1,111,971 students
6,913,128 GPA observations
Discontinuity= 0.02***
SE = 0.004

N = 6,679
8,979 GPA observations
Discontinuity = 0.23***
SE = 0.070

Fig. A1. Final test score distribution first attempts: term 1 vs term 2.

situation A). This is as expected, since for a first attempt in term 1 it is not possible to convert points to grades and concurrently to target grade manipulation effectively. Yet, the right panel (situation B) shows a substantial and significant discontinuity at the passing threshold for first attempts observed in term 2, indicating that a significant proportion of students is transferred from the left to right of the test score distribution. The McCrary density discontinuity result of 0.23 thus indicates that either teachers use their discretion to enable students to pass the subject (and graduate) in the first attempt ($\Delta_i > 0$) and/or that students have (temporarily) boosted their ability to pass the subject. To distinguish whether it is the teacher and/or student creating this difference between situation A and B, the sample is restricted to contain only the high-stakes retaking students (N = 136,638). Whereas high-stakes retaking students represent roughly 12% (136,638 out of 1,118,650) of the total sample, they make up 15% (1,031 out of 6,679) of the students observed to have at least one first attempt in term 2. This corroborates the aforementioned selective nature of this group. The reason they are selected for the results displayed in Figure 7 is that if a subject exam's first attempt is observed in term 2 for this subsample of students, it by definition is a high stakes event (i.e. given that they need a retake for graduation, failing this subject for which the first attempt is observed in term 2 will be detrimental for their propensity to graduate).

The upper left McCrary density test results relate to the subject-specific first-attempt GPA distribution observed in term 1 and the observed negative discontinuity indicates that high-stakes retaker are more frequently (just) failing a particular subject. Yet, when for these students the GPA distribution is analyzed for the subject(s) for which a first attempt is observed in term 2, a markedly different picture emerges in that now a positive discontinuity is observed.

This discontinuity is larger than for the overall population of students for which first attempts are observed in term 2 (i.e. 0.35 versus 0.23 overall), which is in line with the high-stakes nature of this subpopulation of students. Furthermore, the bottom panels in Fig. A2 indicate that results

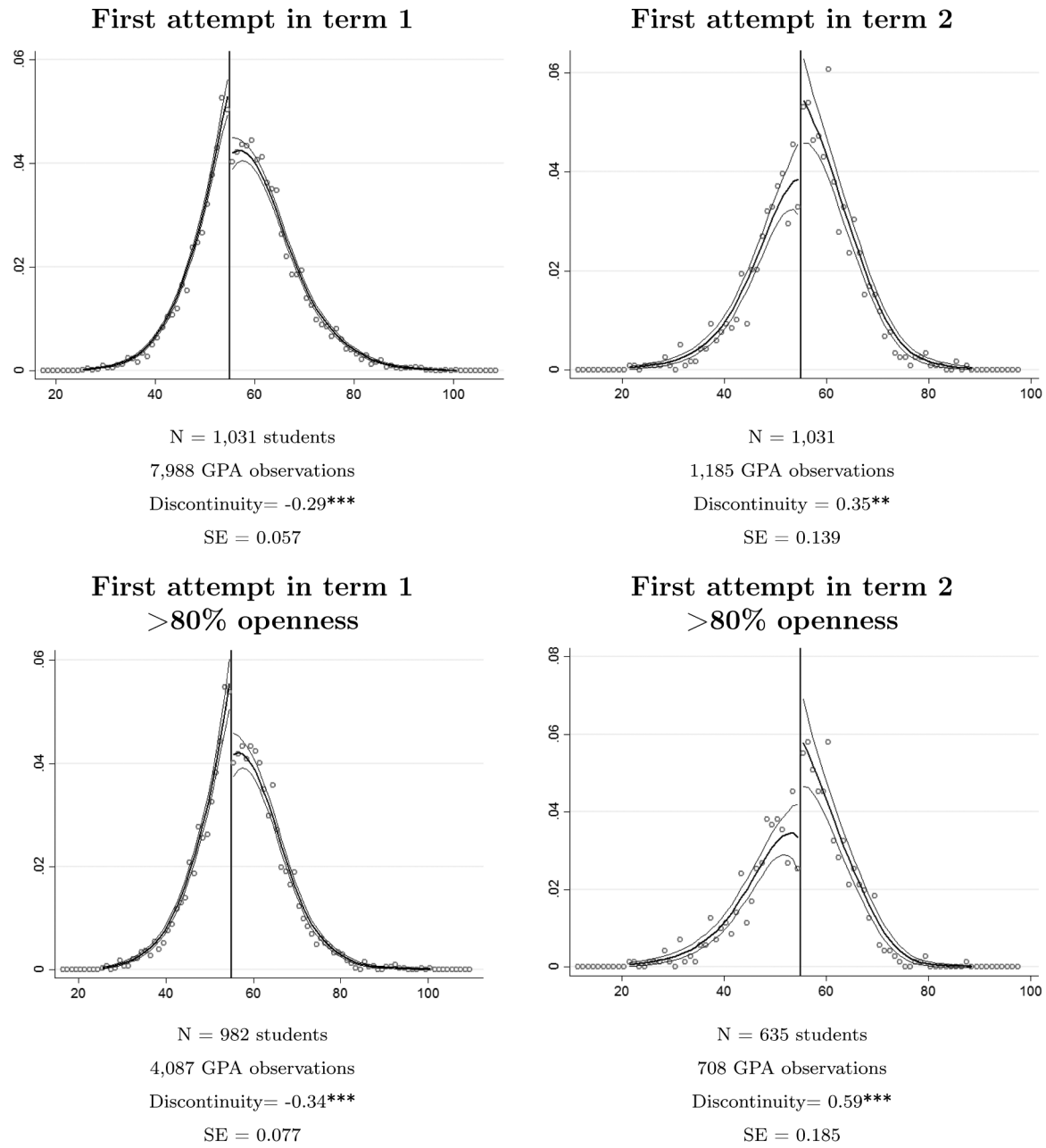


Fig. A2. Final test score distribution of high-stakes retakers' first attempt: term 1 vs term 2.

for exams with more than 80% open questions are similar to other exams in terms of their first attempt term 1 distributions, but even larger positive discontinuities are observed when it concerns a first attempt exam observed in term 2 (i.e. 0.59 versus 0.35 overall). While these results only concern a small subsample of the overall student population, it reaffirms that teachers exploit their discretion to artificially improve performance as to let students graduate who would otherwise not have graduated. Furthermore, this phenomenon is observed in a context when potential mean reversion does not come into play, reveals itself when the *N*-factor is known, is larger in magnitude when students are at risk of not graduating, and larger when teachers have more discretion when grading the exam.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.econedurev.2019.07.002 .

References

Burgess, S., & Greaves, E. (2013). Test scores, subjective assessment, and stereotyping of ethnic minorities. *Journal of Labor Economics, 31*(3), 535–576.

Dee, T. S., Dobbie, W., Jacob, B. A., & Rockoff, J. (2016). *The causes and consequences of test score manipulation: evidence from the new york regents examinationsTechnical Report.* National Bureau of Economic Research.

Diamond, R., & Persson, P. (2016). *The long-term consequences of teacher discretion in grading of high-stakes testsTechnical Report.* National Bureau of Economic Research.

Hanna, R. N., & Linden, L. L. (2012). Discrimination in grading. *American Economic Journal: Economic Policy, 4*(4), 146–168.

Jacob, B. A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago public schools. *Journal of Public Economics, 89*(5), 761–796.

Kuhlemeier, H., & Kremers, E. (2012). *De praktijk van de eerste en tweede correctie van het CSE. Verslag van een landelijke enquêteTechnical Report.* Arnhem: Cito.

Kuhlemeier, H., & Kremers, E. (2013). *De Praktijk van eerte en tweede correctie. Samenvatting van onderzoek naar het functioneren van het CSETechnical Report.* Arnhem: Cito.

Lavy, V. (2008). Do gender stereotypes reduce girls' or boys' human capital outcomes? evidence from a natural experiment. *Journal of Public Economics, 92*(10), 2083–2105.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics, 142*(2), 698–714. https://doi.org/10.1016/j.jeconom.2007.05.005.

McMillan, J. H., & Nash, S. (2000). Teacher classroom assessment and grading practices decision making.

Neal, D. (2013). The consequences of using one assessment system to pursue two objectives. *Journal of Economic Education, 44*(4), 339–352.

Schuurs, U., Kuhlemeier, H., & Gitsels, H. (2017). De invloed van het lt-examenverslag op de scores. *Levende Talen Tijdschrift, 18*(4), 25–35.